2024 年度名古屋大学学生論文コンテスト

奨励賞

大規模言語モデルを用いた青春の定量的な定義法の考案

情報学部 3 年　樹神 宇徳

大規模言語モデルを用いた青春の定量的な定義法の考案

1. はじめに

青春という言葉は現在日本で広く使われる言葉であり、青春小説、青春期、青春群像劇など様々な語彙に使用される言葉である。しかし、一般的に流布している語彙であるのにも関わらず、青春という言葉の一般的な定義は曖昧であり、使用者やコンテキストによってその意味を大きく変える。これは、個人によって思い浮かべる青春の情景が異なる事に起因すると考えられる。それによって「青春」を客観的な指標で評価する手法が存在しておらず、定量的に青春度を示すことが困難であった。そこで本論文では、青春度を客観的な指標で定量化する手法を考案する事によって本問題を解決する。本論文では依岡(2013)が主張する、

ドイツの小説・演劇が、急速な近代化の中で日本の若者が抱いたアイデンティティの不安と制度の間で揺れる心情に触れ、 共感を呼び、青春という概念に独特の色をもたらしたといえる。また一方では、近代化で 失われていった「ふるさと」の代用として、精神の拠り所とされた[1]。

という論を支持し、依岡(2013)の「青春」の脱俗性と純潔さを本論文における青春の基礎的な定義として話を進めていく。「脱俗性」とは俗世間の気風から抜け出ることを表す言葉である。つまり、脱俗性がある状態というのは一般の考え方や行動から逸脱している(意外である)ことを指している。これに加えて、「純潔さ」とはけがれがなく心が清らかなことを示す言葉である。この純潔さは、Graham ら(2013)が提唱する「道徳基盤理論」における Sanctity(神聖さ)/Purity(純潔さ)の基盤に根付いている概念であると考えられる[2]。Cameron ら(2015)は、道徳内容のタイプとそれによって生じる情動は一対一対応となっており、Purity(純潔さ)が損なわれると Disgust(嫌悪)の情動が呼び起こされると主張している[3]。このことから純潔さは、嫌悪という「ネガティブ」な概念の対極の概念であると解釈することができ、逆説的に「ポジティブ」に受け入れられる概念であると考えられる。以上の議論から本論文では、「意外性」と「ポジティブさ」を定量的に測定することによって「青春」を定量的に定義する手法を模索する。一方、依岡(2013)は、「一方で、青春は明るくポジティブなものばかりでもなかった。」と主張しており、青春のポジティブさについては議論の余地がある[1]。しかし本論文では青春の定量的な定義を主目的とするため、「純潔さ」の定義を簡素化し「ポジティブさ」として読み替えることにする。

1.1 大規模言語モデル

近年、OpenAI 社が開発した ChatGPT を筆頭に大規模言語モデルが社会の注目を集めている。大規模言語モデルとは、膨大な量のテキストデータから学習された言語の確率モデルのことであり、ある時点での単語を今まで予測した単語群を用いて予測する人工知能の事を指す[4]。つまり、これらのモデルは事前に学習されたテキストデータを元にして、最も次に来そうな単語を今まで予測した文章から予測している。このことから、多様で巨大なテキストデータによって学習された大規模言語モデルは人間の考え方や判断の確率モデルと考える事が出来る[5]。よって本論文では大規模言語モデルを用いて文章の生成確率から、文章が示す内容の意外性を判定する。これは前章で述べた「青春」の脱俗性に対応すると考える事が出来る。また大規模言語モデルはファインチューニングによって、特定のドメインに於いての文章の分類タスクにおいても適用が出来る[6]。ファインチューニングとは、事前学習したモデルの重みを新しいデータで訓練する手法であり、特に特定のタスクにおけるモデルの性能を向上させるために、そのタスクに必要なドメイン知識を持つデータを追加することによってモデルの特定のタスクにおける性能を向上させることが出来る[7]。このことは、文章のネガティブ・ポジティブを判定する BERT モデルをファインチューニングすることによって作成することが出来る可能性を示している。この判定結果は前章で述べた「青春」の純潔さに対応すると考える事が出来る。以上の大規模言語モデルを用いて定量化した二つの指標を使うことによって、「青春」を文章から定量的に測定する事が出来る。

1.2 情報エントロピー

情報エントロピーとは、情報理論の概念であり、ある事象が起きた際にそれがどれほど起こりにくいかを表す尺度である[8]。数学的には、事象 E が起こる確率を P(E)とするとき、事象 E が起こる情報エントロピーI(E)は、

$$I(E) = -\log_2 P(E)$$

と表される。情報エントロピーは高ければ高いほどその事象が起こらない確率は高くなり、直感的には情報エントロピーはその事象が起きた時の意外性を示している。よって、情報エントロピーを用いることによって意外性を定量的に示すことが出来る。

2 方法論

本実験では、Google 社が開発した大規模言語モデルである Bidirectional Encoder Representations from Transformers(BERT)を用いる。BERT は予測したい単語の出現確率を文章の左にある単語群と右にある単語群の位置関係から予測しているモデル

であり、ChatGPT などに代表される Generative Pre-trained Transformer(GPT)とは異なり文の前後関係から単語の出現確率を予測することが出来る[9]。GPT は予測したい単語の出現確率を文章の左にある単語群のみから予測するモデルである[10]。即ち、GPT は予測したい単語以降の文章を考慮せずに単語の出現確率を計算する。以上の特性からマスクされた文章を事前分布とし予測された後の文章を事後分布とすることができるため、予測された文章がどれほど意外かを測定することに関して BERT は GPT より本実験に適している。本実験では、文章の意外さを計算する為に東北大学自然言語研究グループが提供している日本語事前学習済み BERT モデル(tohoku-nlp/bert-base-japanese)を用いる[11]。現在多くの大規模言語モデルは英語を中心としたテキストデータによって学習されており日本語の文章の出力に対する理解が不足しているため、これに対応するために日本語で学習されたモデルを使用することは日本語のテキストを用いたタスクに於いてより適切である[12]。また、文章がネガティブかポジティブかを判定する手法に於いては、Takami(2022)が作成した日本語の感情分析に特化させた BERT モデル(koheiduck/bert-japanese-finetuned-sentiment)を使用する[13]。本モデルは与えられた文章を positive, neutral, negative の確率を合計が 1 になるように出力する。本モデルは他の BERT モデルやルールベースの手法よりも複数のベンチマークにおいて高い精度を出すことが出来る[14]。以上の二つだけでは、非文(文章として成り立っていない文)が入力された際に、その生成確率の低さから不当に高い文章の意外性を示す可能性がある。そこで非文が生成されたことを検知するために、Koki(2023)が作成した日本語の流暢度を判定する BERT モデル(liwii/fluency-score-classification-ja)を使用する[15]。このモデルは与えられた日本語の文章が流暢である確率を 0 から 1 までの範囲で出力するモデルである。このモデルは小川ら(2020)が作成した日本語文法誤りデータセットにおいて分類の精度を表す指標である AUC が 0.9811 を記録したことから、文法的に間違った文章を高い精度で判定できると考えられる[16]。

2.1 青春情報エントロピーの定義

以上の三つの BERT モデルを用いて、定量化された青春度である青春情報エントロピー$I_{adolescence}$を以下のように定義する。

$$I_{adolescence}(S) = -\log_2 P_{unusual}(S) - \log_2(1 - P_{positive}(S)) - \log_2(1 - P_{fluency}(S))$$

ただし、S は文章、$P_{unusual}$は文章が意外である確率、$P_{positive}$は文章がポジティブである確率、$P_{fluency}$は文章が文法的に正しい確率を表している。文章が意外である確率$P_{unusual}$は以下のように定義する。

$$P_{unusual}(S) = \left(\prod_{i=1}^{|S|} P_{word\_probability}(i, S)\right)^{\frac{1}{|S|}}$$

ただし、$P_{word_probability}(i,S)$は文章Sの$i$番目の単語(トークン)をMASKトークンにしたときに文章Sのi番目の単語が予測される確率、$|S|$は文章Sに含まれる単語(トークン)の数を示している。つまり$P_{unusual}$は文章に含まれている単語を一つ選び隠し、隠された単語が予測される確率の幾何平均である。このことから、$P_{unusual}$は与えられた文章の一単語の平均予測確率になっている。また青春情報エントロピーの項である、$-\log_2 P_{unusual}(S)$は文章が意外である情報エントロピー、$-\log_2(1-P_{positive}(S))$は文章がポジティブと判定されない情報エントロピー、$-\log_2(1-P_{fluency}(S))$は文章が文法的に間違った文章と判定される情報エントロピーである。情報エントロピーは確率が低くなると高くなるので、文章のポジティブ度が上がれば青春情報エントロピーも向上させる為に、文章がポジティブに判定されない確率$1-P_{positive}$の情報エントロピーを青春情報エントロピーは含む。同様の理由から、文章が文法的に間違っている確率$1-P_{fluency}$の情報エントロピーも青春情報エントロピーは含む。情報エントロピーはその確率が排反である時に限り二つの情報エントロピーを足すことが出来る。$P_{unusual}$ $1-P_{positive}$ $1-P_{fluency}$は文章Sに依存せず明らかに排反であるため、青春情報エントロピーはこれら3つの確率の情報エントロピーの総和で表されている。

2.2 指標の妥当性の判定法

前章で提案した青春情報エントロピーが実際に文章の青春度を表しているかを確認するために、評価用の指標を作成する。まず、評価用のデータを作成するために以下の手順でデータを作成する。

1.「私はAをする。」と言う文章のAに1から10のランダムな個数MASKトークンを挿入する。

2. MASKトークンが挿入された文章のMASKトークンを東北大学自然言語研究グループの提供するBERTモデルで、文章の開始に近い場所のMASKトークンから予測される単語をBERTの確率分布に従いMASKトークンと置換する。

3. 文章にMASKトークンが生成されなくなるまで手続き2を繰り返す。文章にMASKトークンが含まれなくなったら文章を保存する。

4. 保存された文章の種類が1000種類になるまで手続き1に戻って繰り返す。

以上の手続きを行う事によって評価用の文章データ1000個を作成する。その後、評価用データの青春情報エントロピーが高い 40 個を取り出す。ここで異常に青春情報エントロピーが高い上位4つのデータを除外し、36個のデータとする。次に作成した文章データから正答の順位データを作成する。正答の順位データの作成するにあたり、ELYZA.incの提供する大規模言語モデルであるLlama-3-ELYZA-JP-8Bを使用する[17]。

Llama-3-ELYZA-JP-8B は日本語の生成能力に関するベンチマーク評価である ELYZA-tasks-100 において 80 億パラメータという軽量のモデルでありながら、OpenAI 社の開発した GPT-3.5 Turbo モデルを上回る性能を達成したモデルである[18]。このことから、Llama-3-ELYZA-JP-8B は日本語を扱う能力が非常に高く本タスクにおいて一般的な「青春」の感覚を適切に判定するのに優れていると考えられる。本論文では Llama-3-ELYZA-JP-8B を用いてシェフェの一対比較法(浦の変法)で正答データを作成する。シェフェの一対比較法(浦の変法)は、複数人の評価者に全ての選択肢の一対比較を提示しそこから選択肢の平均嗜好度を算出する方法である[19]。この方法では、比較順序を考慮する。つまり、一対の試料を提示する際に、時間的な順序や空間的な位置の効果が、データに影響すると思われる時は、この方法を用いることによってそれらの影響を考慮した平均嗜好度を計測することが出来る。本実験では複数人の評価者を再現するために、Llama-3-ELYZA-JP-8B のシード値を変更することで擬似的に複数人が評価しているかのようにする。シード値とは計算機が疑似乱数を生成する為に使用される設定値であり、その疑似乱数の生成の再現性を担保する事が出来る。本実験では Llama-3-ELYZA-JP-8B が出力する単語の確率を元に、疑似乱数を用いてどの単語が予測されるかを文頭から文末まで選んでいく。本実験では評価者を 2 人として実験を行う。以上の議論から以下の手順で正答の順位データを作成する。

1. Llama-3-ELYZA-JP-8B のシード値を変更し固定する。
2. 評価用文章データから二つの文章を選択する。
3. 二つの文章をプロンプト

   「以下の二つの文章の内、どちらの方がより青春的である文章と思うかを答えてください。

   A:(一つ目の文章)

   B:(二つ目の文章)

   回答は必ず A、B のどちらかで回答してください。」

   を入力しどちらの方がより青春を感じる文章かを判定する。
4. 評価用文章データから全ての二つの文章の取り出す組み合わせが行われるまで手順 2 に戻って繰り返す。
5. 予め決めていた評価者の数だけ手順 1 に戻って繰り返す。
6. 一対比較データからシェフェの一対比較法(浦の変法)を用いて平均嗜好度を計算する。

以上の青春情報エントロピーと正答データの平均嗜好度の相関係数を出すことで、青春情報エントロピーの有用性を検証する。

3. 実験結果

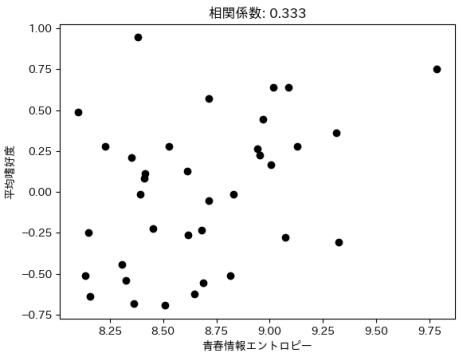

図 1 実験結果のプロット

青春情報エントロピーと正答データの平均嗜好度を計算しプロットした図 1 に示す。図 1 は x 軸に青春情報エントロピーをプロットし、y 軸に平均嗜好度をプロットしている。図 1 から青春情報エントロピーと平均嗜好度の相関係数は 0.333 である。ここで、青春情報エントロピーと正答データの平均嗜好度の相関に有意性があるかを、帰無仮説と対立仮説を以下のように設定する事によって確認する。

帰無仮説H_0:実験結果に相関がない。

対立仮説H_1:実験結果に相関がある。

帰無仮説H_0であると仮定すると、青春情報エントロピーと正答データの平均嗜好度は独立した分布に従うと考えられることから、検定統計量$t = |r|\sqrt{n-2}/\sqrt{1-r^2}$は自由度$n-2$の t 分布に従う。ただし、n は試料数である 36、r は相関係数である 0.333 である。よって統計検定量 t=2.180 は自由度 34 の t 分布に従うことから、危険度 p=0.035(<0.05)より有意水準 5%で棄却され、対立仮説が採択される。以上より二変数青春情報エントロピーと正答データの平均嗜好度には有意に相関があることが示された。

相関係数	自由度	t 値	p 値
0.333	34	2.18	0.035*

*p<0.05

表 1 二変数分布の統計検定量

4. 考察

以上の実験結果から、青春情報エントロピーと平均嗜好度の間に弱い正の相関がある事が確認された。この結果は本論文が主張している、青春情報エントロピーが「青春さ」を表す指標であるという主張を支持する結果である。本実験では弱い正の相関しか得られなかったが、使用するモデルの性能の向上によってより高い相関係数を得られることが期待される。このことから本論文の提案する青春情報エントロピーには青春を客観的に定量化する手法として一定の妥当性を持っていると考えられる。しかし、本実験では比較的短い文章での検証しか行っておらず、長い文章でも同じ妥当性を持つかについては今後の実験で明らかにする必要がある。また、「青春」の意味を定

義する先行研究の少なさから青春の多様的な意味を拾いきれずに単純化してしまっている。よって、今後の「青春」に関する研究の発展によって新たな項が青春情報エントロピーに追加されより精度が上がる可能性がある。青春情報エントロピーは考案されたばかりの指標であり、今後の研究で様々な発展系が生まれることが期待できる。また、本実験で考案した意外性やポジティブさなどを情報エントロピーとして解釈する手法は「青春」の定義にとどまらず、他の多様的な意味をもつ定量化の難しい概念にも応用が可能であると考えられる。他の概念に対しての本手法の有用性については今後の実験で確かめられるべきである。

5. 結論

複数の懸念点はあるものの、本実験では青春度を客観的な指標で定量化する手法である青春情報エントロピーを考案し、その有効性を検証することが出来た。これによって、定量的に青春度を示すことが困難であるという問題を解決する事が出来た。今後の展望として、他の多様的な意味をもつ定量化の難しい概念への応用や長文に対する本指標の有用性の検証が挙げられる。

6. データへのアクセス

本実験で使用したコードとデータは Github にて下記のレポジトリで公開している。
https://github.com/kotama7/seisyun_information_entropy

［参考文献］

[1] 依岡隆児(2013)『旧制高校からみた「青春」概念の形成』国際日本文化研究センター 44巻。

[2] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, Peter H. Ditto(2013) *"Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism"* Advances in Experimental Social Psychology Volume 47 Pages 55-130 https://doi.org/10.1016/B978-0-12-407236-7.00002-4.

[3] C. Daryl Cameron, Kristen A. Lindquist, and Kurt Gray(2015) *"A Constructionist Review of Morality and Emotions: No Evidence for Specific Links Between Moral Content and Discrete Emotions"* Personality and Social Psychology Review 19(4) 1-24

[4] 持橋大地(2022)『大規模言語モデル（LLM）とロボティクス』日本ロボット学会誌 40巻10号。

[5] Michael Franke, Polina Tsvilodub, Fausto Carcassi(2024) *"Bayesian Statistical Modeling with Predictors from LLMs"* https://doi.org/10.48550/arXiv.2406.09012.

[6] 杉崎 光一, 阿部 雅人, 全 邦釘(2023)『大規模言語モデルの専門領域への適用に関する検討』AI・データサイエンス論文集 4巻3号。

[7] Saket Dingliwal, Ashish Shenoy, Sravan Bodapati, Ankur Gandhe, Ravi Teja Gadde, Katrin Kirchhoff(2022) *"Prompt Tuning GPT-2 language model for parameter-efficient domain adaptation of ASR systems"* https://arxiv.org/abs/2112.08718.

[8] Claude E. Shannon(1948) *"A Mathematical Theory of Communication"* The Bell System Technical Journal Vol. 27 pp 379-423 623-656.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova(2019) *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"* https://doi.org/10.48550/arXiv.1810.04805.

[10] Tianzheng Troy Wang(2021) *"GPT: Origin, Theory, Application, and Future"* ASCS CIS498/EAS499 Project and Thesis.

[11] 東北大学自然言語研究グループ(2024) *"bert-base-japanese"* アクセス日:2024-12-30 https://huggingface.co/tohoku-nlp/bert-base-japanese.

[12] Kaito Sugimoto(2024) *"Exploring Open Large Language Models for the Japanese Language: A Practical Guide"* https://doi.org/10.51094/jxiv.682.

[13] Takami Kouhei(2022) *"bert-japanese-finetuned-sentiment"* アクセス日:2024-12-23 https://huggingface.co/koheiduck/bert-japanese-finetuned-sentiment.

[14] Yifan Sun, Hirofumi Tsuruta, Masaya Kumagai, Ken Kurosaki(2024) *"Topic Modeling and Sentiment Analysis on Japanese Online Media's Coverage of Nuclear Energy"* https://arxiv.org/html/2411.18383v1.

[15] Koki Ryu(2023) *"fluency-score-classification-ja"* アクセス日:2024-12-30 https://huggingface.co/liwii/fluency-score-classification-ja.

[16] 小川耀一朗, 山本和英(2020) 『日本語文法誤り訂正における誤り傾向を考慮した擬似誤り生成』 言語処理学会 第26回年次大会 発表論文集。

[17] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, Akira Sasaki(2024) *"Llama-3-ELYZA-JP-8B"* アクセス日:2025-1-2 https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B.

[18] Akira Sasaki and Masato Hirakawa and Shintaro Horie and Tomoaki Nakamura(2023) *"ELYZA-tasks-100: 日本語instructionモデル評価データセット"* アクセス日:2025-1-2 https://huggingface.co/elyza/ELYZA-tasks-100.

[19] 浦 昭二(1956) 『一対比較実験の解析 日本科学技術連盟官能検査研究会資料』 1-8。